2010 International Conference on Cyberworlds

Measuring Multimodal Synchrony for Human-Computer Interaction

Dennis Reidsma, Anton Nijholt
Human Media Interaction
Faculty EEMCS, University of Twente
Enschede, The Netherlands
{dennisr,anijholt}@cs.utwente.nl

Wolfgang Tschacher, Fabian Ramseyer
University Hospital of Psychiatry
University of Bern
Bern, Switzerland
{tschacher,ramseyer}@spk.unibe.ch

***Abstract*—Nonverbal synchrony is an important and natural element in human-human interaction. It can also play various roles in human-computer interaction. In particular this is the case in the interaction between humans and the virtual humans that inhabit our cyberworlds. Virtual humans need to adapt their behavior to the behavior of their human interaction partners in order to maintain a natural and continuous interaction synchronization. This paper surveys approaches to modeling synchronization and applications where this modeling is important. Apart from presenting this framework, we also present a quantitative method for measuring the level of nonverbal synchrony in an interaction and observations on future research that allows embedding such methods in models of interaction behavior of virtual humans.**

Keywords - Synchrony, nonverbal communication, measurement, virtual humans, HCI

I. INTRODUCTION

Synchrony is an important element in non-verbal human-human interaction. It has a role as one of the social signals that help building – and maintaining – a relation between people, be it for the duration of the interaction, or longer term. As such, it may also be a useful indicator of someone's attitude towards the interaction, or their interaction partner.

Currently there is a growing amount of attention for this and other kinds of social signaling in research in Human-Computer Interaction (HCI). Sensors and sensing algorithms are developed for recognizing social signals; the recognition results are used to analyze and/or support human-human interaction; and research is carried out to build systems that can use similar social signaling devices in human-computer interaction, both with and without using humanoids such as virtual humans and social robots as interaction metaphor.

In this paper, we present the following regarding nonverbal synchrony as social signaling device. First we define what kind nonverbal synchrony in human-human communication this paper concerns. Next, we give a short overview of various roles that research into synchrony plays in the field of HCI. Finally, we present a method to quantitatively measure the level of nonverbal motion synchrony in human-human conversations, and discuss how this method must be extended in order to be able to use it in the HCI contexts that we presented.

II. NONVERBAL SYNCHRONY IN HUMAN-HUMAN INTERACTION

People adapt to each other in interaction. They mirror each others' gestures and postures; they converge in their choice of vocabulary; they do not always speak at the same time that another person wants to speak; and show their mutual adaptation in many other ways. By now, this behavior is commonly known under the general terms entrainment and synchrony, although it is known under many other names as well, often with slightly different meanings (Nagaoka et al. [1]). This paper specifically concerns the dynamic, temporal, form of synchrony: the extent to which people adapt the timing and coordination of their behavior to each other.

Literature such as the work of Crown [2], Ramseyer and Tschacher [3, 4] or Nagaoka et al. [1] on interactional synchrony or coordinated interpersonal timing in communication present experimental results that suggest a relation between, on the one hand, being able to coordinate one's actions in an anticipatory manner to those of one's interlocutor, and, on the other and, a positive evaluation of the conversation partner and of the (effectiveness of) the interaction. Crown [2], for example, relates interpersonal timing to affective relation in dyads, and concludes that a 'like/dislike/-unacquainted' condition has a strong relation with interpersonal timing. Ramseyer and Tschacher [4] performed a large scale quantitative study of synchrony in a psychotherapeutic context; using the analysis method developed by Ramseyer [5] they found a clear correlation between synchrony and certain positive therapy outcomes in a data set of 125 therapy sessions by 80 dyads. Nagaoka et al. [1] discuss experiments with rhythmic entrainment (meaning, in their case, convergence of latencies between utterance and response, for speakerA/¬ListenerB and speakerB/-listenerA transitions), showing how dynamics and alignment are a important elements of synchrony tendency. Which, in turn, as they argue, is important for conveying rapport and empathy, promotion of understanding emotion and making yourself assessed positively by the other.

III. NONVERBAL SYNCHRONY AND HUMAN-COMPUTER INTERACTION

Computational detection of synchrony in human-human interaction

Ambient Intelligence is one HCI context in which a lot of importance is placed on the detection and interpretation of all facets of the interaction between humans [6]. As discussed before, nonverbal synchrony has strong correlations with various aspects pertaining to the quality and smoothness of an interaction. Automatically measuring synchrony could therefore be of great use in, for example, smart meeting

978-0-7695-4215-7/10 $26.00 © 2010 IEEE
DOI 10.1109/CW.2010.21

rooms, in which the interaction between inhabitants of the environment is analyzed for real-time support and for retrieval. Starting from the detection of periods of higher or lower synchrony, one could try to find points of 'good' interaction, or automatically annotate attitude-related phenomena such as agreement, disagreement, argumentation, and decision points.

Teleconferencing and synchrony

Teleconferencing is a situation in which the flow of social signals between people is notoriously impoverished. It is an active topic of HCI research to improve teleconferencing applications in such a way as to restore some of the communication that is disrupted by this. One of the many social signals that get lost – and perhaps on of the most difficult to restore, given the time delay factors always involved in long-distance teleconferencing – is synchrony or entrainment (McGrath, [7]).

This problem has been addressed specifically by Watanabe et al. [8, 9]. They propose, among other things, a remote communication system that uses avatars to represent remote partners, instead of video. The virtual avatar embodiment of the remote partner, displayed to the local partner, allows the system – by appropriate animation of the avatar – to show respiration and other physiological and bodily behaviors as they were measured from the remote partner, in a clearer and more understandable way than would be possible using video. Their idea is that synchrony will occur more easily when enough involuntary relevant bodily behavior is displayed. However, in that situation one still has to deal with the time lag caused by the long-distance network link. This might have a negative impact on the occurrence of synchrony, too. This issue is addressed in their proposal for a "Speech-driven embodied interaction system". There, they still use an avatar representation of the remote partner, which is then animated in appropriate ways to display synchrony and other listening behavior. In this case, however, the appropriate behavior is not determined by measuring this behavior on the remote person and displaying it on the avatar, but by locally calculating what would be the best nonverbal behavior to display in (synchrony) relation to the speech of the local participant. This might increase the fluency of the conversation, but displaying such behavior when it does not mirror any underlying real synchrony between the participants of course raises important issues of ethics, trust and believability in teleconferencing, which the authors do not address at all.

Synchrony for entertainment computing

As there are so many positive affective aspects associated with synchrony, it is perhaps unsurprising that some researchers have recently worked to explicitly harness synchrony effects (the rhythmic, temporal synchrony that we address in this paper) for entertainment computing. For example, Weinberg and Driscoll [10] and Crick et al. [11] built robotic drummers that interacts with human co-musicians or a conductor in rhythmic dimensions. Their focus is on the musical expressiveness and the collaborative music making. Michalowski et al. [12] built a robot that dances rhythmically in response to movements of the user, synchronizing to his or her rhythm. They developed it with a strong focus on interpersonal coordination and interactional synchrony, for research into research, therapy, and entertainment (Kozima et al. [13]).

Tanaka and Suzuki [14] explicitly modeled two-way entrainment for their dancing robot Qrio in order to achieve more engagement from the user. A core concept in their interaction model is the repetition of sympathy and variation to it: "We consider that there are two concepts that should be basic units designing the interaction model. One is sympathy between the two (human and robot),and the other is variation to it. Our basic hypothesis is that by the repetition of the sympathy and variation, long-term interaction can be realized." Finally, Tomida et al. [15] attempt to achieve 'entertaining interaction' between two humans by trying to elicit entrainment quite directly. In their MiXer installation, the authors aim to elicit synchronization between two human users of the system: the "rhythms" of a user are sensed using bio sensors (heart rate, perspiration, etc) and the rhythms are presented to the other user using "bio feedback" (visualisation of those rhythms). The other user may then tap a button synchronized to the biofeedback display; the assumption is that an engaging type of "entrainment" between the users will occur in response to this process.

Synchrony in natural interaction with Virtual Humans and Social Robots

More specifically in the context of interaction between humans and Virtual Humans(VH),we do know that at least to a certain extent interactional synchrony also works for human-VH interaction. For example, Suzuki et al. [16], working on prosody, say that echoic humming mimicry has a positive influence on affective perception of the conversational partner, even if that partner is a computer. Bailenson and Yee [17] also specifically address the dynamics of the movement: interactional synchrony, in the form of mimicry (repeat head movements of partner after 4 secs) is effective for Virtual Humans to be more persuasive and effective.

This is all not very surprising, as Reeves and Nass [18] already showed that this type of aspects in human-human communication transfer to human-media communication. More on the topic of timing, Robins et al. [19], working with robots rather than virtual humans, conclude qualitatively from an exploratory study about "Rhythm, kinesics, body motion and timing" that "[...] responding with appropriate timing so as to mesh with the timing of human actions encourages sustained interaction" and "Robot-human temporal interaction kinesics will eventually need to be studied deeply in order to put this dimension within the purview of HRI designers".

Lately, several research groups have started to address such synchrony-like temporal interaction for Virtual Humans. Gratch et al. [20] considered the role of the proper timing of nonverbal feedback of a virtual human, for creating a feeling of rapport between user and agent. Reidsma et al. [21] built a Virtual Conductor that interactively conducts an ensemble of human musicians. The focus of their system is

on the temporal interaction of leading and following. The music being conducted and played serves as a kind of medium to steer the interaction. Nevertheless, the exact timing is not enforced externally; the avatar simultaneously adapts to the timing displayed by the users and attempts to get the users (musicians) to adapt to its timing, and thus in a sense can be said to implement temporal interactional synchrony.

IV. A METHOD FOR MEASURING NONVERBAL SYNCHRONY

A prerequisite for almost any of the HCI contexts sketched above is a method for measuring the amount of non-verbal synchrony in interaction. Between two humans in an interaction or between a user and a computer system (VH or other); to make the system adapt itself, show a reaction in a way that is contingent on the level of synchrony, or to evaluate the quality of the interaction.

Wilson and Wilson [22] and Varni et al. [23] describe computational models for entrainment processes – but such are not necessarily suitable for automatic measurement of synchrony. The method of assessment by expert observers, or by analyzing self-report questionnaires, such as used by Ramseyer and Tschacher, and Tschacher et al. [5, 24] can be a powerful device, but most of the applications rather need an automatic, quantitative, method.

There have also been more than a few attempts at finding a rhythmic organization in speech that can be calculated automatically. If this were possible, one could use it for detecting synchrony through looking at these rhythms of the speech of the conversation partners. However, it is not trivial to find such a rhythmic organization [25, 26]. Cummins [27] was able to detect synchronization between people in a specific task in which the people had to read out a text together synchronously. He measured the amount of synchrony by looking at the lag between the speakers on specific points of the text, or by performing a dynamic time warping algorithm on the spoken sentences – where the amount of warp is more or less inverse to the amount of synchrony – but this approach is not generalizable enough as it only works when you have a specific task in which two people have to perform the same content at the same time (e.g., synchronous speaking or playing a piece of music).

Watanabe et al. [9] calculate the maximum value of the windowed cross-correlation between the head movements of the participants (determined using tracking devices attached to the head). They use this to show that their avatar-based teleconferencing system leads to more synchrony between speakers, as the maximum value of the cross-correlation over the whole interaction is significant higher when they have the avatars display the respiration of the remote person.

Here we describe in more detail a method for automatically calculating synchrony developed by Ramseyer [3], which combines and extends a few separate approaches. Two video streams are recorded from the interaction between two persons. In both videos the amount of movement by the person in the video is measured as a function of time, by an image difference computation [28]. Next, the time-lagged cross-correlation between the two movement functions is computed to determine if the people move synchronously [29]. Since they might not move exactly at the same time, but with a short delay in response to each other, this computation is done many times, comparing each window of person A with all windows of person B that are within a lag of -2..+2 secs. This leads to an image much like the leftmost graph in figure1. In that graph, the white 'highlights' are periods where a high level of motion synchrony is detected between the persons. However, it is conceivable that the measured synchrony would result as a matter of chance from the comparison of any two moving people, whether they are in interaction or not. To compensate for this possibility of 'chance synchrony', Ramseyer introduced 'randomly shuffled pseudo-interactions' where the data of one of the two persons is shuffled around in blocks covering larger periods [5].

Calculating the cross-correlations on this shuffled pseudo-interaction still compares the movements of the same two persons, but now at two different moments – it does not represent a real interaction anymore. The result of this process is show in the right-most graph in figure 1. Comparing these two calculations shows whether the measured level of synchrony is higher than would be expected by chance. Ramseyer and Tschacher [4] validated this method on data taken from psychotherapy sessions, showing that the calculations have a significant correlation with other measures of synchrony or quality of interaction and outcome of the therapy.

V. TOWARDS AN EXTENSION OF THE METHOD FOR USE IN HUMAN-COMPUTER INTERACTION

Before we look at how this method could be applied in HCI, two questions need to be asked. The first question is: "To what phenomena is this method sensitive?". Or, more informally: what does it actually measure? Not just 'movement occurring at the same time'. The movements should also be performed with similar timing. Or to phrase it differently: when a person leans forward en back again, the amount of movement varies from 'low movement' (sitting), through 'high' (starting to go forward), 'low' (being at the most forward-leaning position), and 'high' (going back again) back to 'low' (sitting still again).

This is a graph with two peaks, which would correlate well with a similar movement pattern of two peaks from the interlocutor, if the timing of these peaks is similar. Note, that the peaks need not be timed regularly (metronome-like). If both interlocutors performed fast forward lean, then slowly went back, it would also correlate well, and less so if one of them did it the other way around (slow forward, fast back). The second question is "What is the relation between synchrony and quality of interaction?" Synchrony is context dependent, and not necessarily "the more the better".

Too much synchrony and mirroring makes people uncomfortable, not only in therapy but in any kind of interaction. Boker and Rotondo [31] explicitly say that it is the breaking of synchrony as much as the formation of synchrony (symmetry) that makes the process of communication work well.

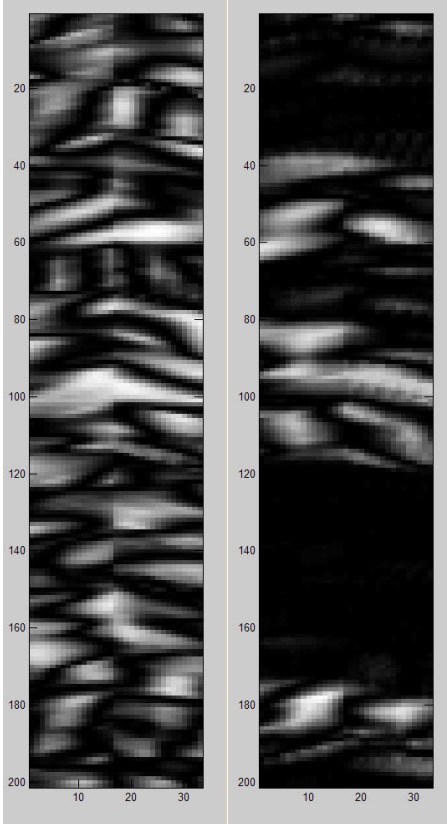

Figure 1. Cross-correlations of movements by two persons in a conversation, generated from a fragment of 200 windows (8 sec) in a free conversation between two people. Vertical: time in interaction. Horizontal: lag(-2..2 sec). Left shows the cross-correlations of the actual interaction; right shows the cross-correlation of the pseudo-interaction in which the movement data of one person has been shuffled around. This particular image has been reproduced from [30], in which the method by Ramseyer and Tschacher was followed.

If we want to look at patterns of forming and breaking synchrony between people, we need to move from the global view of the interaction taken by Ramseyer et al. to a more local view. It is not enough to say that the overall interaction displayed significant synchrony. Instead of just looking at an overall measure for whole conversation, we must look if we can relate specific patches of high synchrony in the interaction to specific conversational events in the interaction. Also, we want to be able to say "here in the interaction, and here, there was an above average high level of synchrony" and then draw conclusions about the quality of interaction at these points in the interaction.

If we know what kinds of interactional behavior lead to "white patches in the graph", do we then conversely also know that "white patches in the graph" mean that meaningful interactional synchrony has occurred? Clearly not. After all, consider the rightmost graph in figure 1. It also shows "white patches in the graph", and there is no interaction at all between the behaviors being compared there. It is not enough to calculate a local aggregate measure of the correlations, such as average, highest peak, or size of the "white patches", when the pseudo-interaction shows similar patches of high cross-correlation at different time lags. So what is needed is a metric that determines that a certain patch of high correlation could not have been caused by chance, similar to what the method of Ramseyer and Tschacher determines for the overall interaction.

Without such a local adaptation of the metric, we can only use the cross-correlations to say something about the overall amount of synchrony in a longer-term interaction. This then, defines a major task for future development of the method described above.

ACKNOWLEDGEMENTS

The authors would like to thank the people appearing in the various data recordings that have been used for the paper. This work has been supported by funding from the EU-FP7 project SEMAINE, the COST Action 2102 (Cross-Modal Analysis of Verbal and Nonverbal Communication), and the EU-FP7 Network of Excellence SSPNet. This paper only reflects the authors views and funding agencies are not liable for any use that may be made of the information contained herein.

REFERENCES

[1] C. Nagaoka, M. Komori, and S. Yoshikawa, "Synchrony tendency: interactional synchrony and congruence of nonverbal behavior in social interaction," in Proceedings of the 2005 International Conference on Active Media Technology, 2005. (AMT 2005)., 2005, pp. 529–534.

[2] C. L. Crown, "Coordinated Interpersonal Timing of Vision and Voice as a Function of Interpersonal Attraction," Journal of Language and Social Psychology, vol. 10, no. 1, pp. 29–46, 1991. [Online]. Available: http://jls.sagepub.com/cgi/content/abstract/10/1/29

[3] F. Ramseyer, "Synchronisation nonverbaler Interaktion in der Psychotherapie [Nonverbal synchrony in psychotherapy]," Ph.D. dissertation, Bern, 2008.

[4] F. Ramseyer and W. Tschacher, "Synchrony in dyadic psychotherapy sessions," in Simultaneity: Temporal Structures and Observer Perspectives, S. Vrobel, O.E. Rossler, and T. Marks-Tarlow, Eds. Singapore: World Scientific, 2008, ch. 18, pp. 329–347.

[5] F. Ramseyer and W. Tschacher, "Synchrony: A core concept for a constructivist approach to psychotherapy," Constructivism in the Human Sciences, vol. 11, no. 1, pp. 150–171, 2006.

[6] R. J. Rienks, A. Nijholt, and D. Reidsma, Meetings and Meeting Support in Ambient Intelligence, ser. Mobile communication series. Norwood, MA, USA: Artech House, 2006, ch. 17, pp. 359–378.

[7] E. McGrath, "Time matters in groups," in Intellectual Teamwork: Social & Technological Foundations of Cooperative Work, J. Galegher, R. E. Kraut, and C. Egido, Eds. Hillsdale, N.J.: Laurence Earlbaum, Assoc., 1990.

[8] T. Watanabe, "E-cosmic: embodied communication system for mind connection," Robot and Human Interactive Communication, 2004. ROMAN 2004. 13th IEEE International Workshop on, pp. 1–6, Sep. 2004.

[9] T. Watanabe, M. Ogikubo, and Y. Ishii, "Visualization of respiration in the embodied virtual communication system and its evaluation," International Journal of Human-Computer Interaction, vol. 17, no. 1, pp. 89–102, Mar. 2004.

[10] G. Weinberg and S. Driscoll, "Robot-human interaction with an anthropomorphic percussionist," in CHI '06: Proceedings

of the SIGCHI conference on Human Factors in computing systems. New York, NY, USA:ACM, 2006,pp. 1229–1232.

[11] C. Crick, M. Munz, and B. Scassellati, "Synchronization in social tasks: Robotic drumming," in Robot and Human Interactive Communication, 2006. ROMAN 2006. The 15th IEEE International Symposium on, Sep. 2006, pp. 97–102.

[12] M. P. Michalowski, S. Sabanovic, and H. Kozima, "A dancing robot for rhythmic social interaction," in HRI '07: Proceedings of the ACM/IEEE international conference on Human-robot interaction. New York, NY, USA:ACM, 2007, pp. 89–96.

[13] H. Kozima, M. Michalowski, and C. Nakagawa, "Keepon – a playful robot for research, therapy, and entertainment," International Journal of Social Robotics, vol. 1, no. 1, pp. 3–18, Jan. 2009.

[14] F. Tanaka and H. Suzuki, "Dance interaction with QRIO: a case study for non-boring interaction by using an entrainment ensemble model," in Proceedings of the IEEE International Workshop on Robot and Human Interactive Communication(RO-MAN'04). IEEE Computer Society, Sep. 2004, pp. 419–424.

[15] T. Tomida, A. Ishihara, A. Ueki, Y. Tomari, K. Fukushima, and M. Inakage, "MiXer: the communication entertainment content by using "entrainment phenomenon" and "bio-feedback", "in Advances in Computer Entertainment Technology, M. Inakage, N. Lee, M. Tscheligi, R. Bernhaupt, and S. Natkin, Eds. ACM, 2007, pp. 286–287.

[16] N. Suzuki, Y. Takeuchi, K. Ishii, and M. Okada, "Effects of echoic mimicry using hummed sounds on human-computer interaction," Speech Commun., vol. 40, no. 4, pp. 559–573, 2003.

[17] J.N. Bailenson and N. Yee, "Digital chameleons: Automatic assimilation of nonverbal gestures in immersive virtual environments," Psychological Science, vol. 16, no. 10, pp. 814–819, Oct. 2005.

[18] B. Reeves and C. Nass, The media equation: how people treat computers, television, and new media like real people and places. New York, NY, USA: Cambridge University Press, 1996.

[19] B. Robins, K. Dautenhahn, C. L. Nehaniv, N. A. Mirza, D. Francois, and L. Olsson, "Sustaining interaction dynamics and engagement in dyadic child-robot interaction kinesics: Lessons learnt from an exploratory study," in Proc. of the 14th IEEE International Workshop on Robot and Human Interactive Communication, RO-MAN2005, 2005, pp. 716–722.

[20] Gratch, N. Wang, J. Gerten, E. Fast, and R. Duffy, "Creating rapport with virtual agents," in IVA, ser. Lecture Notes in Computer Science, C. Pelachaud, J.-C. Martin, E. André, G. Chollet, K. Karpouzis, and D. Pelé, Eds., vol. 4722. Springer, 2007, pp. 125–138.

[21] D. Reidsma, A. Nijholt, and P. Bos, "Temporal interaction between an artificial orchestra conductor and human musicians," Computers in Entertainment, vol. 6, no. 4, pp. 1–22, Dec. 2008.

[22] M. Wilson and T.P. Wilson, "An oscillator model of the timing of turn-taking," Psychonomic Bulletin & Review, vol. 12, no. 6, pp. 957–968, Dec. 2005.

[23] G. Varni, A. Camurri, P. Coletta, and G. Volpe, "Emotional entrainment in music performance," in Proceedings of 8th IEEE International Conference on Automatic Face & Gesture Recognition, FG 2008, Sep. 2008, pp. 1–5.

[24] W. Tschacher, F. Ramseyer, and K. Grawe, "Der ordnungseffekt im psychotherapieprozess," Zeitschrift fuer Klinische Psychologie und Psychotherapie, vol. 36, no. 1, pp. 18–25, Jan. 2007.

[25] M. Bull, "An analysis of between-speaker intervals," in Proceedings 1996 of the Edinburgh Postgraduate Conference in Linguistics and Applied Linguistics, May 1996, pp. 18–27.

[26] E. Keller, "Beats for individual timing variation," in Proceedings workshop on The Fundamentals of Verbal and Non-verbal Communication and the Biometrical Issue, ser. NATO Security through Science: Human and Societal Dynamics, A. Esposito, E. Keller, M. Marinaro, and M. Bratanic, Eds.,vol.18. Amsterdam, The Netherlands: IOS Press, May 2007, pp. 115–128.

[27] F. Cummins, "Measuring synchronization among speakers reading together," in Proc. ISCA Workshop on Experimental Linguistics, 2006, pp. 105–108.

[28] K. Grammer, R. Honda, A. Schmitt, and A. Jütte, "Fuzziness of nonverbal courtship communication unblurred by motion energy detection," Journal of Personality and Social Psychology, vol. 77, no. 3, pp. 487–508, 1999.

[29] S. M. Boker, M. Xu, J. L. Rotondo, and K. King, "Windowed cross-correlation and peak picking for the analysis of variability in the association between behavioral time series." Psychological Methods, vol. 7, no. 3, pp. 338–355, Sep. 2002.

[30] A. Nijholt, D. Reidsma, H.van Welbergen, H.J.A. op den Akker, and Z. M. Ruttkay, "Mutually coordinated anticipatory multi-modal interaction," in Nonverbal Features of Human-Human and Human-Machine Interaction, Lecture Notes in Computer Science, vol. 5042. Berlin: Springer Verlag, 2008, pp. 70–89.

[31] S. M. Boker and J. L. Rotondo, "Symmetry building and symmetry breaking in synchronized movement," in Mirror Neurons and the Evolution of Brain and Language, M. Stamenov and V. Gallese, Eds., 2003.